

Weekly report

1 Done

1.1 RelationLine 软文

听说你也想化身福尔摩斯？RelationLine 了解一下

从一份数据中找到蛛丝马迹已经不是一件难事。然而，在解决真实问题的时候，可以利用的数据往往不止一份。当线索一条条摆在面前，要如何从中整理出唯一的真相呢？近日，阿里巴巴-浙江大学前沿技术联合研究中心发表了最新研究成果 **RelationLine**——一个可以从异构城市数据中推理出个人关系网络的可视分析系统。

不同的城市数据可以为个人关系网络的构建提供不同的线索。比如，轨迹数据可以揭示不同的伴随关系；POI（感兴趣的地点）数据统计了不同区域内商场、学校等功能区的分布；社交媒体数据可以提供已发生事件的文字描述。基于轨迹数据可以形成初步的关系网络，基于其他信息，我们可以对关系网络进行分析和解释。

综合多种数据推理出个人关系网络，在刑侦等领域有着巨大的应用价值。然而，这也很具有挑战。特别是轨迹数据的分析，由于数据获取途径多，覆盖范围大，如果可以有效利用，将对构建个人关系网络起到重要的作用。然而轨迹数据由于数据来源不同，精度不同，存在着难以匹配的问题。此外，不同时间、地点的同行或伴随关系往往代表着不同的关系种类。如何找到不同关系人也是一大挑战。

针对这两大挑战，浙江大学 CAD&CG 实验室陈为老师团队和阿里云城市大脑团队合作完成了一个可视分析系统 **RelationLine**。针对第一个挑战，**RelationLine** 给出了一种基于四叉树的轨迹匹配方法。该方法首先将地图上的点通过四叉树编码成一长串数字。在匹配的过程中，用户可以通过设定匹配位数来调整匹配的精确度，以适应不同需求。接下来，根据数据的不同，**RelationLine** 提供两种计算轨迹之间相似度的方法。对于签到、打卡等只记录了几个轨迹点的数据，**RelationLine** 统计匹配的轨迹点占总轨迹点个数的百分比；对于比较完整的轨迹数据，**RelationLine** 计算匹配时长占总轨迹时长的百分比。这样一来，不同精度、格式的数据都可以进行相互匹配。

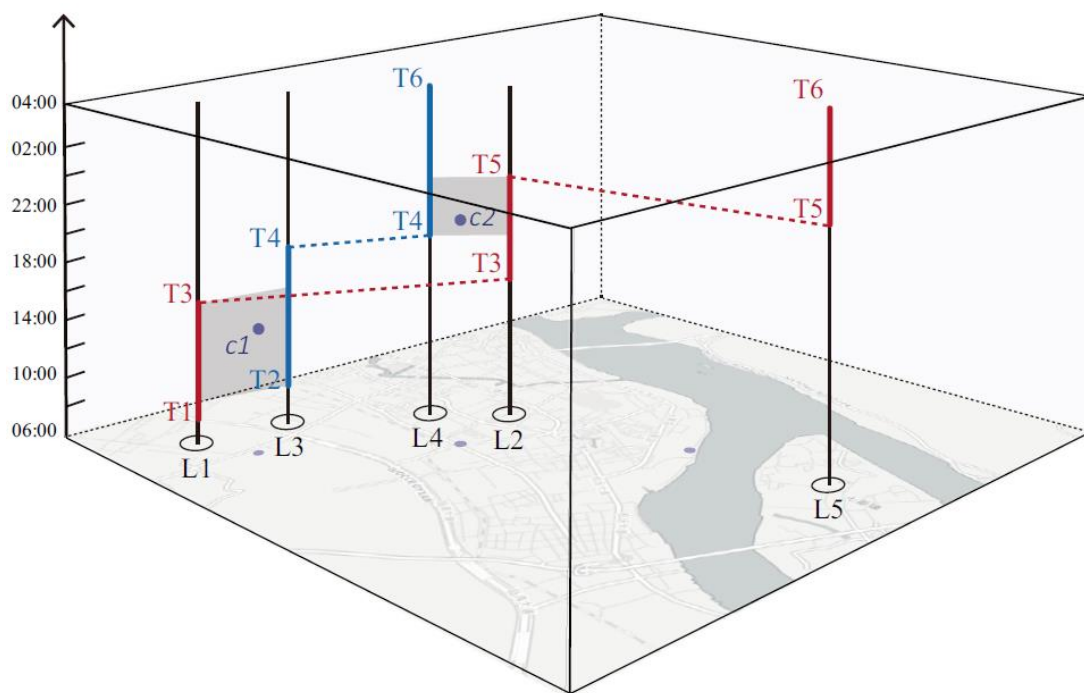


图 1 轨迹的时空匹配方法

针对的第二个挑战，RelationLine 给出了一套基于探测过滤的方法（如图 2），以辅助用户从杂乱无章的信息中找到有用的线索，并把它们串联起来。RelationLine 的探测过滤方案从识别目标（移动电话或出租车）开始。在探测阶段，分析人员需要通过轨迹匹配提取一些满足要求的候选人，这些候选人可能会和目标一起停留在某个地点，或者在特定的时间间隔内跟随目标发生了一系列的位置变换。为了选取可以最快找到合适候选人的轨迹段，分析人员需要观察目标的轨迹并利用其他数据（如 POI 和社交媒体树）来推断目标的生活模式。在探测阶段找到的候选人可能与目标有明确的关系。在过滤阶段，我们需要通过渐进式轨迹匹配来进一步探索目标的个人关系网络。用户可以迭代地调整不同的过滤条件（不同轨迹段的轨迹相似度），并重新确定自我中心关系。用户可以提取最相关的个体并保存在以自我为中心的网络中并根据伴随关系进行进一步研究和修复已构建的自我网络。

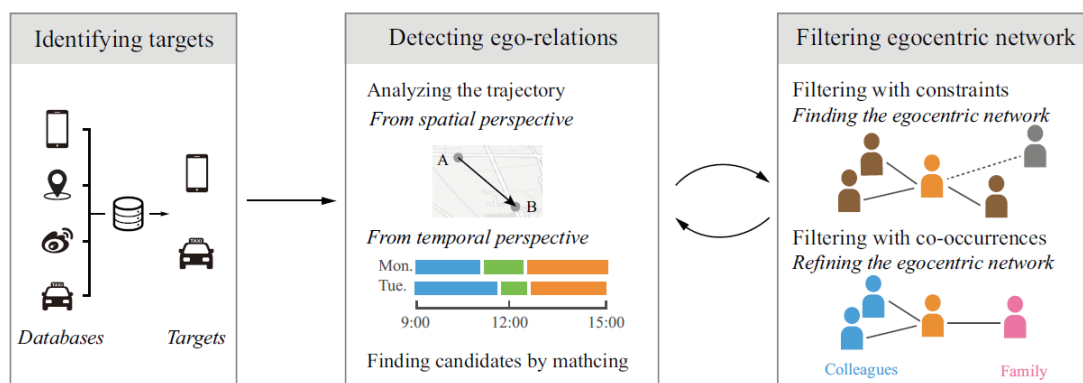


图 2 探测过滤的方法

注 1：论文中数据为实验室数据，非阿里内部或城市大脑的数据。论文 RelationLines: Visual Reasoning of Egocentric Relations from Heterogeneous Urban Data，将发表于 ACM Transactions on Intelligent Systems and Technology，2018，主要作者包括陈为、夏菁和王叙萌等。

1.2 Paper review

1.3 Paper reading

[2017 CHI] ***Trust, but Verify: Optimistic Visualizations of Approximate***

Queries for Exploring Big Data

Authors propose an approach called *optimistic visualization* from the perspective of user experience.

Different from familiar per-group confidence intervals (seeing each group as having its own confidence interval), the distribution uncertainty is a total amount by which the whole visualization is likely to be imprecise. Distribution uncertainty recognizes that uncertainties are not independent. The expression with Euclidean distance is shown below.

$$\sqrt{\sum_{\text{group } i} \left(\frac{\text{group } i\text{'s value}}{\text{total group value}} - \frac{\text{estimated group } i\text{'s value}}{\text{total estimated group value}} \right)^2}$$

Based on distribution uncertainty, the per-group confidence interval can only be computed with worst-case estimates. Though *Sample+Seek*, *optimistic visualization* produces approximate results quickly and computes precise results in the background. Thus, it allows analysts to explore approximate results interactively. Related system--Pangloss implements the approach.

[2017 CHI] ***GraphScape: A Model for Automated Reasoning about***

Visualization Similarity and Sequencing

In this work, authors define a taxonomy of specification edits that captures possible transitions among Vega-Lite unit visualization. The process consists three steps:

1) Identifying Edit Operations

Category	Edit Operations	Description	Example
Mark	$Mark_A_Mark_B$	Change the mark type from A to B , or vice versa.	Change Bar marks to Line marks.
Transform	$Scale$	Add, change or remove an axis scale transform.	Apply a log scale on the x-axis.
	$Sort$	Change the sort order of a visualized data field.	Sort 'Maker' on x-axis in descending order of 'Mean Price'.
	Bin	Discretize values of a visualized field into bins.	Bin 'Price' into 10-unit-wide bins.
	$Aggregate$	Aggregate values according to an aggregation function such as sum, mean or median.	Aggregate 'Price' to mean.
	$Modify\ Filter, Add / Remove\ Filter$	Filter data records according to a filter predicate function.	Filter 'Price' values ≤ 100.0 . Keep 'Maker' values equal to 'Company A' or 'Company B'.
Encoding	$Transpose$	Swaps the (x, y) or (row, column) channels.	Transpose the x- and y-axes.
	$Move$	Move a field on one channel to another channel.	Move 'Price' from x-axis to y-axis.
	$Add / Remove$	Add or remove a field from an encoding channel.	Add 'Price' to x-axis. Remove 'Price' from x-axis.
	$Modify$	Replace a field in a channel with another field.	Replace 'Price' with 'Maker' on the x-axis.

Table 1. Edit operations for transitions among visualization specifications. The table is sorted in ascending order of estimated pairwise cost.

2) Ranking Edit Operations

Authors formulate a set of triplet comparisons: given a source visualization and two target visualizations, which target visualization is easier to interpret in the context of the source?

3) Deriving Edit Operation Costs

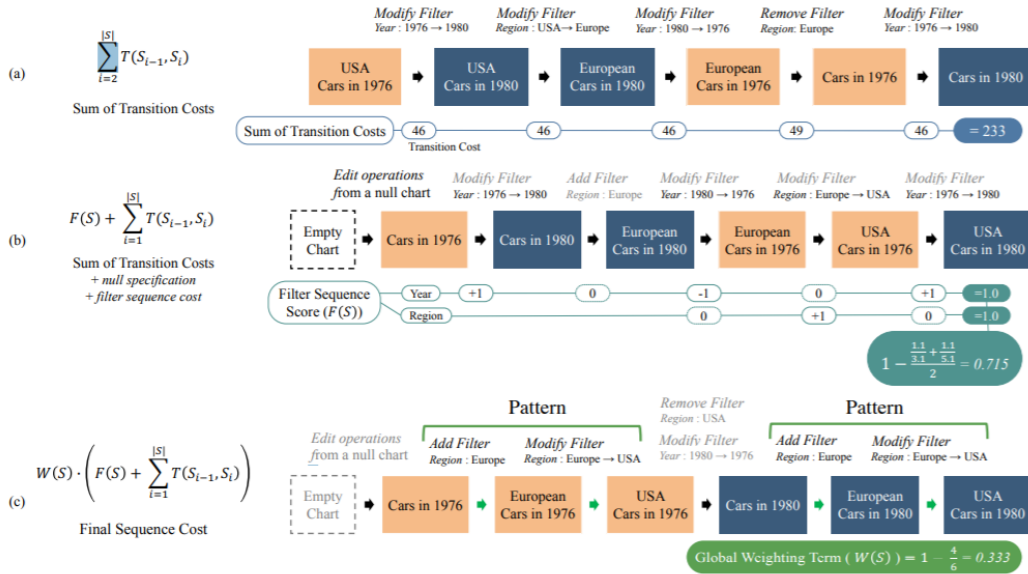


Figure 4. GraphScape sequence cost calculation. (a) Lowest-cost sequence when minimizing the sum of transition costs only. (b) Lowest-cost sequence after adding an initial null specification and filter score. (c) Lowest-cost sequence with a global weighting term to reward consistent subsequences.

The *GraphScape* cost model provides an objective function for determining optimized sequences. It is beneficial to narrative visualization..

2 Progress

Item	Deadline	Current progress	Remark
Patent	-		-
Monography	4.30		-